ISTA 321| Data Mining

Dr. Nicholas DiRienzo
Email: ndirienzo@email.arizona.edu
Online office hours: Generally Thursday or Friday at 11am
Online Zoom office: https://arizona.zoom.us/j/2251199844
Course Homepage: D2L / Slack
SLACK JOIN LINK: https://join.slack.com/t/slack-hkj2733/shared_invite/zt-1mvg6l7kz-4JJNlcnOjlktH4KuyHitCQ
Final Exam: **Friday, March 3rd**

Course Description: This course will introduce students to the theory and practice of data mining for knowledge discovery. This includes methods developed in the fields of statistics, large-scale data analytics, machine learning and artificial intelligence for automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns. Topics include understanding varieties of data, classification, association rule analysis, cluster analysis, and anomaly detection. We will use software packages for data mining, explaining the underlying algorithms and their use and limitations. The course include laboratory exercises, with data mining case studies using data from biological sequences and networks, social networks, linguistics, ecology, geo-spatial applications, marketing and psychology

Course Objectives:
In this course students will:

1. Be introduced to the concepts of supervised and unsupervised learning and how they are used to solve various data-centric problems.
2. Learn the fundamental algorithms of supervised and unsupervised learning (e.g. regression, decision trees, k-means, k nearest neighbors, anomaly detection methods, etc).
3. Learn the critical preprocessing steps and validation methods that are required to ensure such methods are implemented properly.
4. Understand why programming languages are critical for data mining.

Learning Outcomes:
By the end of the semester students will be able to:

1. Define what data mining is as a field and be able to identify how it applies to various use cases.
2. Implement supervised and unsupervised statistical and machine learning methods to answer questions about real world datasets.
3. Articulate the pros and cons of the various methods, models and techniques learned in class and make decisions about which one is best for a given scenario.
4. Describe the technical steps that proceed the application of a given algorithm and justify why they are needed.

Online Course Format:
This course will use several different mediums for learning. Most weekly content will be given in the form of a lesson I have created, R scripts, or chapters from the assigned textbook. I will also post youtube videos that cover and recap content, but these are *not* meant to be stand-alone lessons but rather supplements to the written content provided. Being an online course, you will need to be more self-directed in both your learning and solving problems.

There will be two main coding exercises each week. The first is found in the lessons I provide. You are expected to code through, modify, play with the code in the lessons created to show you how these methods work and relate to the overall goal. The second is the actual **homework** assignments. These will be where you apply the ideas and skills learned in the chapter to a new data set to make inference.

R and Getting Help:
This course will use the programming language R, which all of you should have had at least one course in. The first week is a refresher that will check and make sure you have the skills needed for this class. It's fine if you don't ace it, but if you are *really* struggling then you should chat with me.

One of the key skills to develop when programming is figuring out how to solve your own problems. This can be one of two ways. **First**, google… google is a great and normal way to solve problems, but it's also easy to get bogged down with the thousands of hits one various stack overflow pages. **Second**, if googling your problem doesn't work, **jump over to our Slack channel** (Check out setup tutorial on D2L). Slack is the go-to communication software for tech companies as it's faster and more direct than emailing. It also preserves an archive of communication. So, if you can't solve an issue, go post an informative question (or answer!) to the Slack.

Regarding emailing me for help - essentially, don't do it if it's related to course content! If you are struggling with an issue you should post it to Slack! Chances are someone else is struggling as well, or someone has already solved the issue. I want you all to work together to solve problems. Now, if I see wrong answers or something not being answered, I'll jump in (eventually), but I will be giving credit for Slack participation (both on answering and asking). Where this gets messy is that although I highly encourage you to help each other, **This does not mean you can directly share code associated with an assignment (this is a violation of UA's Code of Academic Integrity).** So, Slack help should be in the form of providing suggestions, similar code samples, or direction as to where to find help (i.e. a website or where in a lesson we covered something). I'm not going to be too harsh on this as I want you all to participate, though. So if you're unsure if you are giving too much code, just post.

A Few Words on Technology and Communication:

1. You will have access to and will be required to retrieve all course materials from the course page in D2L. Training for D2L can be found online at: http://help.d2l.arizona.edu/students.

2. You will need to have R and R Studio installed and functioning week 1. Here's a tutorial.

3. **Again Slack participation is critical!** If you are having a coding issue, first try and solve it on your own. If you're still struggling, then post it to the class Slack channel. Essentially, if you are about to email me with a homework/class/coding question, post it to slack first. I'm not doing this to save me time, but rather because Slack is *the* communication tool for most tech companies, and learning how to use it is really important. Also, answering questions on Slack also contributes to your participation score. **THIS IS A LARGE, REALLY EASY PART OF YOUR GRADE. IF YOU HAVE ISSUES PARTICIPATING ON SLACK TELL ME ASAP.**

Readings:
Weekly lessons will come via R Markdown HTML files. This will be the foundational content along with teaching you a lot of how to code for the class and problems.

We'll supplement my content with the book *Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. You can download a free PDF from the book's website, although I highly recommend buying a copy. It's both easier to work from a paper copy and is a truly excellent book that will give you a solid foundation for machine learning. This will provide some more conceptual depth that I don't have room for in my book.

If you want some extra R help you can check out the book "R for Data Science" by Hadley Wickham and Garret Grolemund. This book covers how to create full data science pipelines in R (more than we'll be doing here) and is available free here: https://r4ds.had.co.nz/.

Complete List of Assignments with Grade Breakdown:

Activity	Total Percent	Unit Percent	Activity & Notes

Weekly assignments	45%	7.5%	There will be 7 total assignments - one will be dropped
Exams (2)	45%	22.5%	
Slack participation / intro	10%	-	Mix of in-class and message board participation. **To get full credit, students must post/interact at least once per assignment unit.**

Grade Distribution:
90-100% = A "exemplary, far beyond reqs/expectations"
80-89% = B "exceeds requirements/expectations"
70-79% = C "meets requirements/expectations"
60-69% = D "falls short of requirements/expectations"
< 60% = E "repeat of course needed"

Course Schedule: Here is the schedule for the class in terms of topics and due dates. The class is divided into 10 main topics of learning with homework assignments due at the end of each topic.

1. Wednesday, January 11th: Say hi on Slack. **Update R!** Make sure you understand course structure.
2. Monday, January 16: Reviewing R, R data structures, and an overview of data mining
   a. Homework due Saturday, January 21
3. Monday, January 23: Diving deep into linear regression + Understanding through graphing
   a. Homework due Saturday, January 28
4. Monday, January 30: Manipulating and preprocessing your data + Features vs. targets
   a. Homework due Saturday, February 4
5. Monday, February 6: Classification via parametric and nonparametric models
   a. Homework due Saturday, February 11
6. **Tuesday, February 14: Midterm**
7. Wednesday, February 15: Cross validation and understanding types of error + For loops
   a. Homework due Saturday, February 18
8. Monday, February 20: Trees and forests
   a. Homework due Saturday, February 25
9. Monday, February 27: Unsupervised Learning - Clustering
   a. Homework due Wednesday, March 1
10. **Friday, March 3: Final exam**

Exam & Final Format:
This course has a single midterm exam and a final exam, both worth the same amount of points and use the same format. For the exams you will be given three hours to complete the online exam. You will be able to take this exam anytime within the day stated on the syllabus. Once you start your exam you must finish, so be sure to block off enough time to complete it. **Both exams are open book/notes, but you can't use google or any other comparable search engine to find answers (doing so will result in a zero for the question at a minimum)**. The exams will include both conceptual questions, applied questions, as well as general coding skills questions. University policy on final examinations can be found here: https://www.registrar.arizona.edu/courses/final-examination-regulations-and-information

Student Accommodations:

It is the University's goal that learning experiences be as accessible as possible. **If you anticipate or experience physical or academic barriers based on disability or pregnancy, please let me know immediately so that we can discuss options.** You are also welcome to contact Disability Resources (520-621-3268) to establish reasonable accommodations. For additional information on Disability Resources and reasonable accommodations, please visit http://drc.arizona.edu/.

Attendance, Due Dates, and Missing Work:

1. **Missed class assignments or exams cannot be made up without a well-documented, verifiable, excuse (for example, a physician's medical excuse).** Indeed, *due dates are firm*, and late work will be accepted only with a verifiable and valid excuse.
2. The UA policy regarding absences for any sincerely held religious belief, observance or practice will be accommodated where reasonable, http://policy.arizona.edu/human-resources/religious-accommodation-policy.
3. Absences pre-approved by the UA Dean of Students (or Dean designee) will be honored. https://deanofstudents.arizona.edu/absences
4. Arriving late and leaving early is extremely disruptive to others in the class. Please avoid this kind of disruption.
5. The UA's policy concerning Class Attendance and Administrative Drops is available at: https://catalog.arizona.edu/policy/class-attendance-participation-and-administrative-drop

Course Conduct and Campus Policies (be familiar with all campus policies):

1. Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of independent effort unless otherwise instructed. Students are expected to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See: https://deanofstudents.arizona.edu/student-rights-responsibilities/academic-integrity
2. It is the University's goal that learning experiences be as accessible as possible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, please let me know immediately so that we can discuss options. You are also welcome to contact Disability Resources (520-621-3268) to establish reasonable accommodations. For additional information on Disability Resources and reasonable accommodations, please visit http://drc.arizona.edu/.
3. The UA Threatening Behavior by Students Policy prohibits threats of physical harm to any member of the University community, including to oneself. See http://policy.arizona.edu/education-and-student-affairs/threatening-behavior-students.
4. All student records will be managed and held confidentially. http://www.registrar.arizona.edu/personal-information/family-educational-rights-and-privacy-act-1974-ferpa?topic=ferpa
5. The University is committed to creating and maintaining an environment free of discrimination; see http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy.
6. Information contained in this syllabus, other than the grade and absence policy, may be subject to change without advance notice as deemed appropriate by the instructor.